DOCUMENT RESUME

ED 295 993 TM 011 848

AUTHOR Juul, Dorthea; Loewy, Erich H.
TITLE The Selection of Critical Errors on a Medical School Certifying Examination.
PUB DATE 87
NOTE 14p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Certification; Clinical Diagnosis; Higher Education; *Licensing Examinations (Professions); *Medical Schools; Medical Students; Multiple Choice Tests; *Scores; *Test Format; Test Items
IDENTIFIERS *Critical Errors; Patient Management

ABSTRACT

This study analyzed the relationship between selecting critical errors (choices that would be dangerous to patients) and conventional test scores on a medical school certifying examination that included three item formats: regular and weighted multiple-choice questions and patient management problems. Data from a Clinical Certifying Examination administered to 279 seniors as the University of Illinois College of Medicine were analyzed. It was found that while there were significant negative correlations between test scores and number of critical errors made across the three different item formats, there were nonetheless students who passed the examination although they made a relatively large number of critical errors. Implications for teaching and testing are discussed. Four tables present data on critical errors. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED295993

The Selection of Critical Errors on a Medical School Certifying Examination

Juul, Dorthea, University of Illinois at Chicago

Loewy, Erich H., University of Illinois College of Medicine at Peoria

3.4 Item and test analysis

ABSTRACT

This study analyzed the relationship between selecting critical errors (choices that would be dangerous to patients) and conventional test scores on a medical school certifying examination that included three item formats: regular and weighted multiple choice questions and patient management problems. It was found that while there were significant negative correlations between test scores and number of critical errors made across the three different item formats, there were nonetheless students who passed the examination although they made a relatively large number of critical errors. Implications for teaching and testing are discussed.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

DORTHEA JUUL

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM011 849

THE SELECTION OF CRITICAL ERRORS ON A MEDICAL SCHOOL CERTIFYING EXAMINATION

Objectives of the Study

There is concern in the academic medical community that some students may meet requirements for medical school graduation and yet may be prone to making dangerous mistakes which could jeopardize their patients' lives or retard recovery. The faculty committee charged with the responsibility for developing a certifying examination for the University of Illinois College of Medicine (UICOM) has, over the years, viewed the graduation of students who make dangerous mistakes on this examination with alarm.

This study was therefore undertaken to answer the following questions: 1) Is there a correlation between overall test performance and the number of critical errors made? 2) Is there consistency in the selection of critical errors across different item formats? 3) Are there examinees who pass the test and yet make a significant number of critical errors? and 4) Are there examinees who fail the examination but who make only a modest number of such errors?

Review of the Literature

Three studies have recently appeared that address this issue. Grosse (1986) reports on a study of the selection of dangerous responses to multiple choice items on the American Board of Orthopaedic Surgery's 1983 and 1984 certifying examinations. The mean number of dangerous options selected was low: 1.4 (S.D. = 1.3) with 31 possible in 1983 (n = 548) and 2.8 (S.D. = 1.9) with 66 possible in 1984 (n = 746). The Pearson correlations between total test scores and number of dangerous options selected exceeded -1.00 for both years after correction for attenuation due to unreliability. Comparisons were made between different examinee subgroups. U.S. and Canadian medical school

graduates selected significantly fewer dangerous responses than did foreign medical school graduates, and examinees taking the test for the first time selected significantly fewer dangerous responses than did repeaters. These results held across both the 1983 and 1984 examinations.

Because of the high negative correlations between total test scores and number of dangerous options selected, Grosse (1986) concludes that, "It seems unlikely that dangerous option scores could contribute any information about candidate ability not already contained in the total test scores." (p. 465) He adds that the study results do provide further evidence of the tests' construct validity. Candidates generally select few dangerous options indicating there is a good match between their preparation and the test content, and high-scoring examinees make fewer such choices than low-scoring examinees.

Mankin, Lloyd, and Rovinelli (1987) also studied the selection of dangerous answers on four multiple choice specialty board examinations administered to 2,713 examinees between 1981 and 1983. A panel of experts identified dangerous answers on 491 out of a total of 903 multiple choice questions (46%). Scores were combined across the four examinations by expressing the conventional percent correct scores in standard deviation units from the mean and the percent dangerous answer scores on a seven-point scale based on the number of dangerous answers selected compared to the total number of items on the test. Arbitrary pass/fail scores were set at 1.5 standard deviations below the mean for the conventional score and at three for the dangerous answer score.

Overall, the examinees chose dangerous responses for 8% of the items. Those who passed on the basis of their percent correct scores chose dangerous

responses for 6% of the items, and those who failed selected 11%. When the pass/fail rate was compared for the two scoring methods, it was found that the percent with a failing conventional score and a passing dangerous answer score was 3.2, and 10.4% had a passing conventional score and a failing dangerous answer score. The authors conclude that scoring dangerous answers may provide useful information about examinees, particularly those close to the pass/fail score.

Slogoff and Hughes (1987) reach a different conclusion based on a study of 2,449 candidates who took the American Board of Anesthesiology's 1983 written examination. A panel of experts identified 29 items with 43 dangerous responses out of 175 multiple choice questions. The 1,036 candidates who passed the test selected a mean of 1.6 (S.D. = 0.3; range = 0-7) dangerous answers, and the 1,413 who failed selected a mean of 3.4 (S.D. = 0.4; range = 0-10), a significant difference ($p < .001$).

Selection of four or more dangerous responses was used to define a potentially dangerous group for further evaluation. There were 92 passing candidates and 631 failing candidates in this group. For the passing group, their "potentially dangerous" status was not confirmed by ratings of residency performance or by performance on the oral examination given subsequent to the written examination (86 of the 92 took this examination). The authors therefore conclude that, ". . .implementation of alternate scoring by the dangerous answer format would be unnecessarily punitive and unwarranted." (p. 630)

In summary, three studies have looked at the selection of dangerous responses on specialty board examinations. Two of these conclude that scoring in this manner is not helpful, one because of high negative correlations with

regular percent correct scores and the other because of lack of supporting evidence from other information sources. The authors of the third study feel that this approach to scoring may be useful, particularly for students close to the pass/fail score.

Instrument and Methods

The data for this study were derived from a recent Clinical Certifying Examination given to 279 seniors at the UICOM. All students are required to pass this two-day examination prior to graduation. The examination consisted of regular multiple choice questions (RMC), weighted multiple choice questions (WMC), patient management problems (PMP), and a data gathering problem (DGP).

The regular multiple choice questions focus on data interpretation and patient management rather than recall of factual information. The weighted multiple choice questions also focus on patient management, but the options are weighted from +8 to -8 depending on their appropriateness or inappropriateness for the care of the patient at that point in time. The PMPs utilize latent image printing technology, and each problem contains several sections corresponding to various stages in the work-up and management of a patient. Like the weighted multiple choice questions, the options are weighted from +8 to -8 depending on their appropriateness/inappropriateness. Data gathering problems are short answer items usually focussing on some aspect of a patient work-up such as ordering diagnostic studies. There was only one short DGP on this examination which did not lend itself to the identification of critical errors, and it was therefore not included in the study.

Critical errors (CEs) on this examination were defined by the faculty committee as options which, if chosen, would be likely to place the patient in

significant jeopardy. Examples of critical errors are performing a lumbar puncture in the presence of papilledema; performing a gastroscopy in the presence of probable unstable angina; and giving quinidine to a patient in atrial fibrillation with a rapid ventricular response prior to controlling the rate.

Critical errors were initially identified by the full committee during its review of materials selected for the examination. If agreement by consensus was not evident, options were not included as constituting critical errors. The questions were given a final review by one of the authors who is a member of the committee for conformity to the definition.

There were 295 RMCs (usually with 5 options), 48 WMCs (usually with 7 options), and 9 PMPs with a total of 580 options. As Table 1 indicates, there was a total of 108 CEs across the three test formats: 55 RMC, 18 WMC, and 35 PMP. The possible number of CEs per examinee was 63 as some items contained more than one CE. There were 21 RMC, 14 WMC, and 28 PMP.

Student scores included a regular multiple choice score, a weighted multiple choice score, a patient management problem score, and a total score. The total score was a weighted combination of the part scores based on amount of examination time. There were three critical error scores for each of the three item formats plus a total number of critical errors across item formats.

The reliabilities for the different parts of the examination were .86 for the RMC (Kuder-Richardson Formula 20), .49 for the weighted multiple choice (Angoff Formula 12), and .70 for the PMPs (Angoff Formula 12).

To further analyze the relationship between total test performance and number of CEs made, the examinees were divide into four subgroups. These were: 1) those scoring below the minimum pass level of 62% (n = 13); 2)

those scoring at or slightly above the MPL (62-65%; n = 23)); 3) those scoring between the mean and one standard deviation below the mean (66-72%; n = 108); and 4) those scoring above the mean (73% and above; n = 135).

Results

The mean and standard deviation of the number of CEs selected by item type and by subgroup appear in Table 2. For the total group the mean number of CEs on the RMCs was 3.04 (S.D. = 1.69); on the WMCs it was 0.77 (S.D. = 0.79); and it was 1.96 (S.D. = 1.48) for the PMPs. The mean number of total CEs across item types was 5.77 (S.D. = 2.66).

The range of CEs was 1 to 18. For the 13 failing students it was 6-18, and for the barely passing group (n = 23) it was 4-12. For the students with the highest scores in the class (n = 12), the number of CEs ranged from 1-5.

There were 14 students with 11 or more CEs (2 standard deviations above the mean). Of these, four failed the total test, and ten passed. None of these ten scored above the mean, however, and four of them were in the barely passing group.

Correlations between the test scores and number of CEs appear in Table 3. There are significant negative correlations between all of the variables. The largest correlations are between RMC score and RMC CEs (-.50); Total score and RMC CEs (-.46); PMP score and PMP CEs (-.59); RMC score and Total CEs (-.53); PMP score and Total CEs (-.53); and total score and Total CEs (-.59).

Table 3 also contains correlations between critical errors on different item formats and correlations between test scores on different item formats. The correlations between critical errors on different item formats were low though significant ($p < .03$): .12 between RMC and WMC, .16 between RMC and PMP, and .12 between WMC and PMP. The correlations between test scores on

different item formats were moderate: .55 between RMC and WMC, .52 between RMC and PMP, and .35 between WMC and PMP ($p < .000$).

In order to further explore the relationship between total test performance, ANOVA was used to compare the four subgroups on the number of CEs selected, and the results appear in Table 4. There were significant differences among the four groups on the number of CEs by item type as well as total number of CEs.

A posteriori comparisons were made using the Scheffe procedure with the significance level set at .05. For RMC, Group 1 made significantly more CEs than Groups 2, 3, and 4, and Groups 2 and 3 made significantly more CEs than Group 4. For WMC, the only significant pairwise comparison was between Groups 3 and 4. Group 3 made significantly more CEs than Group 4. For PMP, Groups 1, 2, and 3 made significantly more CEs than Group 4. For Total CEs, Groups 1, 2, and 3 made significantly more CEs than Group 4, and Group 1 made significantly more CEs than Group 3.

Educational Significance

Overall, the students made approximately six CEs on this examination: three on RMC, one on WMC, and two on PMP. It is interesting to note that they tended to make fewer CEs on the formats with weighted options, relative to the total number possible (RMC = 21; WMC + PMP = 42). The students were aware that penalties were attached to incorrect choices, and perhaps this information made them more cautious.

There were significant negative correlations between test scores and number of critical errors selected across the three different item formats (regular and weighted multiple choice and patient management problems). The correlations between the number of critical errors made on different item

formats were low although significantly different from zero. The number of CEs made by the failing group was at or exceeded the total group mean for all 13 students. Of the 14 examinees who chose a large number of critical errors, only four examinees failed the test.

When the faculty committee reviewed these results, concern was expressed that students are indeed passing the examination who make significant numbers of CEs, indicating that they lack information and/or decision-making skills which might put their patients at serious risk. In addition to extending the study to a second group of examinees and reviewing the clerkship performance of students who make a large number of CEs, other options being considered include weighting CEs more heavily so that selecting a CE has a greater negative effect on the student's overall score than is currently the case and developing a CE subsection with a required minimum pass level for that section. The examination committee has also prepared feedback for the curriculum committees and department chairs to alert them to apparent weaknesses in their instructional programs.

REFERENCES

Grosse, M. D. "Scores Based on Dangerous Responses to Multiple-Choice Items." Evaluation and the Health Professions, 1986 (9), 459-466.

Mankin, H. J.; Lloyd, J. S.; and Rovinelli, R. J. "Pilot Study Using 'Dangerous Answers' as Scoring Technique on Certifying Examinations." Journal of Medical Education, 62 (1987), 621-624.

Slogoff, S. and Hughes, F. P. "Validity of Scoring 'Dangerous Answers' on a Written Certification Examination." Journal of Medical Education, 62 (1987), 625-631.

TABLE 1
NUMBER OF CRITICAL ERRORS ACROSS ITEM TYPE

	Regular Multiple Choice	Weighted Multiple Choice	Patient Management Problems	Total
Total Number of Critical Errors	55	18	35	108
Possible Number of Critical Errors Per Examinee	21	14	28	63

TABLE 2
MEAN AND STANDARD DEVIATION OF CRITICAL ERRORS

	Total Group (n = 279)	Group 1 (n = 13)	Group 2 (n = 23)	Group 3 (n = 108)	Group 4 (n = 135)
Regular Multiple Choice Critical Errors	3.04 (1.69)	5.46 (1.66)	4.00 (1.71)	3.45 (1.60)	2.32 (1.33)
Weighted Multiple Choice Critical Errors	0.77 (0.79)	1.15 (0.99)	1.00 (0.74)	0.90 (0.85)	0.59 (0.68)
Patient Management Problem Critical Errors	1.96 (1.48)	3.23 (2.49)	2.83 (1.50)	2.21 (1.32)	1.49 (1.30)
Total Critical Errors	5.77 (2.66)	9.84 (3.34)	7.83 (2.42)	6.56 (2.30)	4.39 (1.93)

Group 1 = Total test score less than 62%
Group 2 = Total test score 62-65%
Group 3 = Total test score 66-72%
Group 4 = Total test score 73% and above

TABLE 3
CORRELATIONS

Correlations between Critical Errors and Test Scores

	RMC CEs	WMC CEs	PMP CEs	Total CEs
RMC Score	-.50	-.17	-.30	-.53
	p=.000	p=.003	p=.000	p=.000
WMC Score	-.29	-.27	-.13	-.34
	p=.000	p=.000	p=.014	p=.000
PMP Score	-.24	-.18	-.59	-.53
	p=.000	p=.001	p=.000	p=.000
Total Score	-.46	-.22	-.41	-.59
	p=.000	p=.000	p=.000	p=.000

Correlations between Number of Critical Errors on Different Item Formats

	WMC CEs	PMP CEs	Total CEs
RMC CEs	.12	.16	.76
	p=.024	p=.003	p=.000
WMC CEs	---	.12	.44
	---	p=.022	p=.000
PMP CEs	---	---	.70
	---	---	p=.000

Correlations between Test Scores on Different Item Formats

	WMC Score	PMP Score	Total Score
RMC Score	.55	.52	.93
	p=.000	p=.000	p=.000
WMC Score	---	.35	.72
	---	p=.000	p=.000
PMP Score	---	---	.72
	---	---	p=.000

TABLE 4
ANOVA RESULTS

Critical Errors on Regular Multiple Choice Questions

Source	D.F.	Sum of Squares	Mean Squares	F Ratio	F Prob.
Between Groups	3	186.1804	62.0601	28.195	0.000
Within Groups	275	605.2997	2.2011		
Total	278	791.4800			

Critical Errors on Weighted Multiple Choice Questions

Source	D.F.	Sum of Squares	Mean Squares	F Ratio	F Prob.
Between Groups	3	9.5143	3.1714	5.307	0.0014
Within Groups	275	164.3408	0.5976		
Total	278	173.8551			

Critical Errors on Patient Management Problems

Source	D.F.	Sum of Squares	Mean Squares	F Ratio	F Prob.
Between Groups	3	75.1189	25.0396	12.860	0.0000
Within Groups	275	535.4454	1.9471		
Total	278	610.5642			

Total Number of Critical Errors

Source	D.F.	Sum of Squares	Mean Squares	F Ratio	F Prob.
Between Groups	3	637.5836	212.5279	43.821	0.0000
Within Groups	275	1333.7198	4.8499		
Total	278	1971.3025			

DOCUMENT SUMMARY

Document Id: 4292c
Document Name: CE paper
Operator: juul
Author: juul

Comments:

STATISTICS

OPERATION	DATE	TIME	WORKTIME	KEYSTROKES
Created	08/10/87	15:37	2:00	12819
Last Revised	11/15/87	15:04	:30	1394
Last Printed	11/15/87	15:34		
Last Archived	/ /	:	onto Diskette	

Total Pages: 11
Total Lines: 356
Total Worktime: 4:47
Total Keystrokes: 18685

Pages to be printed 11

Notify U09 on system VOL1.